On new emerging concepts of Modeling Petroleum Digital Ecosystems by Multidimensional Data Warehousing and Mining Approaches

Shastri L. Nimmagadda[1,2], Heinz Dreher[1], Muhammad Nawaz[2], and Kamran Laiq[2]

[1]Curtin Business School – Information Systems, Curtin University, Perth, Western Australia,
[2]Schlumberger, East Ahmadi, Kuwait.
e-mail : snimmagadda@slb.com; h.dreher@curtin.edu.au

Abstract— **Petroleum system and its ingredients are narrated for each and every oil and gas field in each and every petroleum-bearing sedimentary basin. A new concept of ecosystem and its digitization are emerging within the generic petroleum system. Significance of this concept is to make connectivity among petroleum systems through attributes of ingredients and their contextualization and specification. Most popularly known ingredients are geological *structure, reservoir, source* and *seal* rocks. Other ingredients involved are in the form of process of these ingredients such as *maturation* (of source rocks) and *migration* and *timing* (of formation of *structure, reservoir and seal* rocks). One can notice the connectivity among primary petroleum system ingredients, through different processes, such as maturation of source rocks and charging capability and migration of hydrocarbons into suitable structural (*structure*) entrapment areas of *reservoir*. Unless the phenomenon of interconnectivity is understood; integration between ingredients and processes in the context of digital representation and visualization, petroleum system existence and its survival cannot be well explained. Its value cannot be added in terms of petroleum accumulations and volumes, unless these phenomena are explicit. Authors propose ontology based data warehousing and data mining technologies, in which, conceptualization and contextualization of multiple data dimensions (petroleum system's ingredients and processes), integration (within data warehouse environment) and data mining of interpretable emerging petroleum digital ecosystems are accomplished. Multidimensional data warehousing and mining facilitate an effective interpretation of petroleum systems, minimizing the ambiguities involved during *structure* and *reservoir* qualifications and quantifications.**

Index Terms— **ontology, petroleum systems, data warehousing, mining and basins data.**

I. INTRODUCTION

Petroleum system is an information system, in which several elements and processes are described. These elements and processes are described most popularly as entities or dimensions, which presumably, intelligently communicate each other. In the context of broader notion of sedimentary basin, integration of these entities or dimensions, fit with notion of data warehousing, with true concept of metadata representation and integration of multiple petroleum systems. In our study, an attempt has been made to acquire multiple petroleum systems from different sedimentary basins [1] and integrate them, using the concept of data warehousing. Logical structuring of multiple data dimensions is a perquisite for an effective data integration that can facilitate the data mining process. Several data mining views drawn from a data warehouse are used for interpreting domain ontologies, such as precise relationships among structure, reservoir, source and seal elements.

A. Petroleum Ecosystems

An ecosystem is a system whose members benefit from each other's participation via symbiotic relationships [3] and [7] (positive sum relationships). In the context of sedimentary basin research, modeling and petroleum systems analysis, narration of several petroleum systems and different oil and gas fields, in each petroleum system, is complex community, but its environment functioning as an ecological unit. More realistically, it is term of millions of data attributes and properties from volume of DBs of multiple basins and their respective petroleum systems all that store in one place.

B. Ontology as a specification mechanism describing petroleum system

A sedimentary basin of formally represented knowledge is based on a *conceptualization* [3]: the structure, reservoir, source and other maturity and migration path process entities that are assumed to exist in some area of interest and the relationships that hold among them. A conceptualization is an abstract, simplified view of the basin that we wish to represent for some purpose. Every knowledge base, knowledge-based system, or knowledge-level agent is committed to some conceptualization, explicitly or implicitly. In our case, investigating sedimentary basin, its associated petroleum systems, interpreted hydrocarbon plays and producing prospects have complex with some known relationships and many unknown hidden relationships. Multiple dimensions or entities are interpreted among these petroleum systems and each dimension has multiple data attributes and or characteristics. Domain knowledge is key in narrating the ontology and when the knowledge of domain is represented in a declarative formalism, the set of dimensions that can be represented is called, in a broad sense digital basin or with sets of digital petroleum systems. Ontology of a petroleum system is developed and represented based on systematic existence, logic and intelligent design of data structures. Relationships are built among structure, reservoir, source rock, seal rock elements and processes such as timing of formation of elements and migration paths with several logical conceptualizations. During development of domain knowledge among elements and processes, classes, relationships, functions and dimensions (super-type and sub-type) associated with sedimentary basin must be well understood. The connectivity and intelligent communication among these dimensions are well described during course of ontological commitments. A dimension commits to ontology if its observable actions are consistent with the definitions and designs of ontology. The idea of ontological commitment is based on the knowledge-

level perspective. The knowledge level is a level of description of the knowledge of a dimension that is independent of symbol-level representation used internally by the dimension. Knowledge is attributed to dimensions by observing their actions, a dimension 'knows" something if it acts as if it had information and is acting rationally to achieve its goals. The actions of dimensions – including knowledge base servers and knowledge based systems – can be seen through a tell and ask functional interface ([3], [6] and [7]), where a client interacts with a dimension by making logical assertions (tell), and posing queries (ask). Logically, a common ontology defines the vocabulary within which queries and assertions are exchanged among dimensions. Ontological commitments are also agreements to use the shared vocabulary (logic) in a coherent and consistent manner. The dimensions sharing a vocabulary need not share a knowledge base; each knows things the other does not, and a dimension that commits to ontology is not required to answer all quires that can be formulated in a shared vocabulary (logic). A commitment to a common ontology is an agreement to consistency, but not completeness, with respect to queries and assertions using the vocabulary (logic) defined in the ontology.

II. ISSUES AND PROBLEM STATEMENT

Exploration of sedimentary basins is better described by means of interpretation of stratigraphic sequence and structural styles of sedimentary rocks. Petroleum system studies describe genetic relationship between pod of source rock and the resulting petroleum. Investigations of plays describe a series of present-day traps, and of prospects, an individual trap and determine whether an individual trap has any economic value and exploitable with available drilling and production facilities. Several relationships exist among these investigations, which are known (relating the structure-with reservoir) and unknown, which may be hidden among several petroleum system elements and processes. Establishment of concept to commission is an important issue in assessing the feasibility and economic value of a petroleum system. Another issue is description of multiple dimensions that make of petroleum systems, existing in a sedimentary basin.
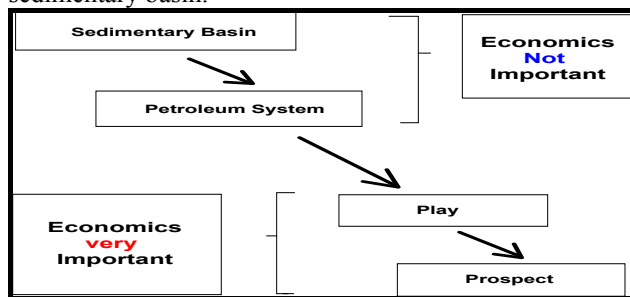

Fig. 1: Four levels of petroleum

In general, narration of petroleum system of any basin is complex and several ingredients, making up the system possess variety of attributes and each attribute is characteristic in its representation, classification and conceptualization. The contextualization is significant, which has been ignored in several knowledge domains of petroleum system. Broadly, four levels of factors are involved in petroleum investigations (Fig. 1). Investigation of sedimentary basin is made categorizing with groups of petroleum systems. Each petroleum system comprises of group of hydro-

carbon fields making up the petroleum system. Each hydrocarbon field is again composed of groups of structural and or stratigraphic plays making up the hydrocarbon field and then finally interpreting the prospect in a particular knowledge domain in terms of its economic leverage. Authors try to build relationships among several basins, petroleum systems, oil/gas fields, hydrocarbon plays and then prospects. These representations and categorizations are ontologically described in different contextualization (conceptualizations) and domains.

A petroleum system is an information system, similar to any information systems described in different domains of applications. Petroleum system requires timely convergence of certain geologic elements, processes and events essential to the formation of the hydrocarbon deposits. As stated earlier, a petroleum basin (or province) is a geologic entity containing at least one or more petroleum systems. The concept of petroleum system is contextualization domain, in which continuous portions of sedimentary volumes may contain pools, describing:
- Reservoirs of similar or dissimilar productive geologic sequence (or seismic sequences)
- Similar or dissimilar chemical compositions
- Similar or dissimilar trap types

As described in Fig. 1, several domain ontologies can be interpreted and genetic relationships built between a particular pod of generating source rock and the resulting petroleum resource. A prospect is conceptual because a successful prospect turns into oil/gas field when drilled or the prospect is unsuccessful, if hydrocarbons are disappeared.
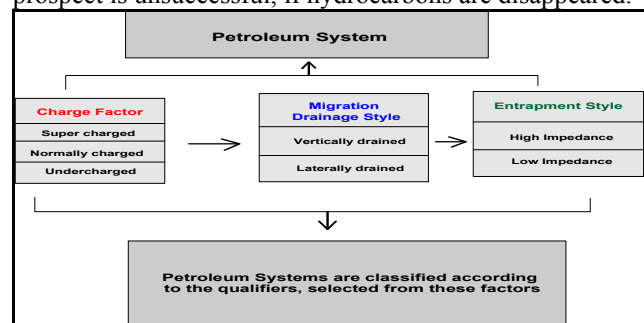

Fig. 2: Genetic classification of petroleum system

Each petroleum system is genetically classified (Fig. 2) in terms of several processing factors, such as charge, migration drainage style, entrapment style. Each charge factor again, is critically categorized in terms of *super-charged*, *normally-charged* and *undercharged*. Under migration process category, hydrocarbons are either vertically and or laterally charged. Under entrapment category, it could be high or low impedance. As a matter fact, authors conceptualize the relationships among these classifications for an effective integration. An effective way of interpreting economically viable petroleum system is to integrate the tectonic framework, sequence stratigraphy, geologic history, thermal history along with sedimentary basin analysis and modeling.

Basin analysis is investigation of a sedimentary basin. Basin modeling is how different ingredients are put together to formulate a hypothesis. The difference between analysis and modeling is that in analysis, an existing item is dissected to determine how it functions, where as in modeling; a hypothetical item is dissected to determine how it should function. Prospect modeling is used on a prospect to

justify drilling, where as a prospect analysis is carried out to investigate why prospect lacked hydrocarbons post drilling.

Several factors such as investigation, economics, geologic time, existence, cost and analysis and modeling are investigated in each sedimentary basins, petroleum system, play and prospect. Hierarchical, relational and networking relationships can be constructed among these factors and levels of investigations. Similarly, elements and processes of petroleum systems are hierarchically and relationally linked among attributes and characteristics of each of these attributes.

III. METHODOLOGIES

The data warehouse approach brings together petroleum systems data from different sedimentary basins of different depositional and geological regimes along with data from other geophysical and geochemical data warehouses. The data warehouse approach is used to benchmark and track the effectiveness of petroleum system productivity over time. It also allows processed (knowledge based) data shared among professionals, geographically distributed worldwide. The need to integrate petroleum systems data from multiple systems and sources is well known [1] and [4]. It is important for data warehouse designers to define the scope, depth, comparability and accuracy of data entering the warehouse. The scope of data refers petroleum systems data, sedimentary basins, data, geological, geophysical and geo-chemical data from multiple periods (time dimension), geographic locations (space dimension). Depth of data refers the level of details. To be comparable, data from multiple dimensions and different sites should adopt the same classifications, as much as possible. No matter how differently data are collected across sites, they are significantly altered for integration before moving into the data warehouse. To reduce the burden of alteration, it is important for petroleum systems analysts and geo-modelers to use compatible software systems to acquire and send data to the repositories. It is also important to standardize the data collection processes. Accuracy of data is desired for all types of data in any given situation and this is fundamental requirement for reliable use of data.

Several data models are deduced that represent ontology, narrating relationships among petroleum system ingredients. Data structures or schemas are modeled in different star-schemas. In these data schemas, there are multiple dimensions narrated and interpreted conceptually with relationships among several dimensions (Fig. 3) and fact data tables, physically.
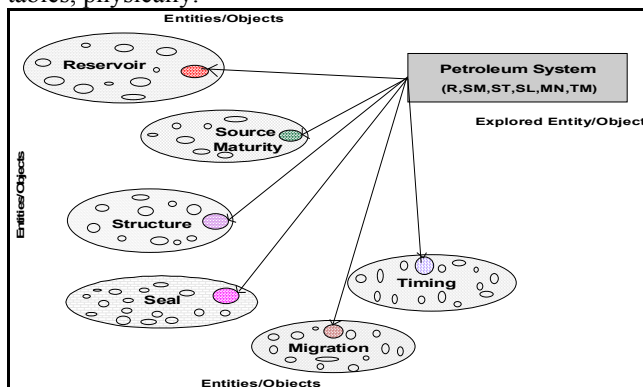

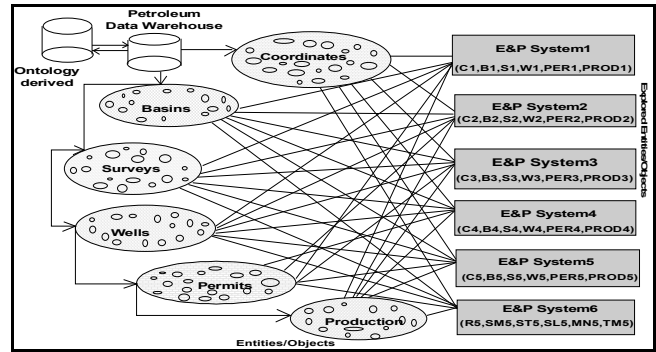Fig. 3a: A model representing relationships of a petroleum system

Fig. 3b: A model representing relationships of multiple petroleum systems

It is important to design intelligently among these multiple dimensions in away to understand or increase domain knowledge that is unknown. For example, the knowledge built among fields, API, reservoir quality, production rate, is significant understanding knowledge of particular petroleum systems, producing for long period of time. Ontology models are constructed for shallow marine and deltaic petroleum systems, in which several dimensions are described, connecting the factual data with one-to-many relationships. These models are represented in star-schema.

IV. DATA MODELING APPROACHES

The data schemas should be simple and flexible enough to modify as per the geological situations and users requirements. If the data schemas are too large, complexity is a significant barrier to widespread adoption of the warehouse technology, because users will find the schema so difficult to understand that they will be unable to write queries and application programs. Schemas must be evolved to grow and support new data types. Limiting the scalability to more data sources though has definite advantage, but in petroleum exploration industry, data structures are very large size.

More logical data structuring is to define single common tables for information that is common to many warehouse data types, to decrease the schema complexity. For example, for many basins and petroleum systems, there may be common reservoirs, structural styles, source and seal types. Each can be implemented through single common table. To this extent, as associative dimension or entity needs to be defined among common structure, reservoir, source, seal and depositional systems. If not single, data warehouse, dimension types used in different spaces of dimension identifiers, uniquely defining that association would require that schema to encode both, *dimension ID* and *dimension type*, since dimensions of different types could have the same ID. Thus a simpler approach is used where by all Data warehouse dimensions share a single space of dimension identifiers within a warehouse instance. The identifiers are known as warehouse identifiers and are integers that are assigned at DB load time. The schemas designed for a warehouse, must have a support with concurrent presence, accessibility and addressability of multiple datasets, multiple versions of a given DB ([3], [8], [9], within a petroleum systems warehouse instance.

The warehouse schema should facilitate coercion of different sources of the same type of data into a common semantic framework. If there is no common information of attribute or data property among data dimensions, then different DBs exist side by side within a warehouse environ-

ment, without having merged. Once common data property information exists, DBs get merged. Several star-schema models drawn in Figs. 4-7, may be merged with logical key attributes among the participating dimensions in the structuring process.
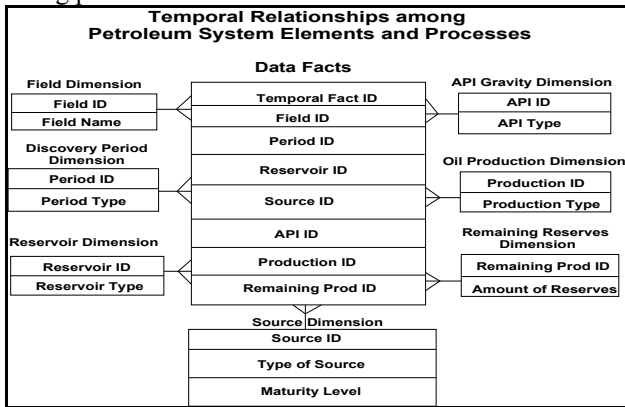

Fig. 4: A star-schema model representing relationships among source rock, API, production, period dimensions and their data facts
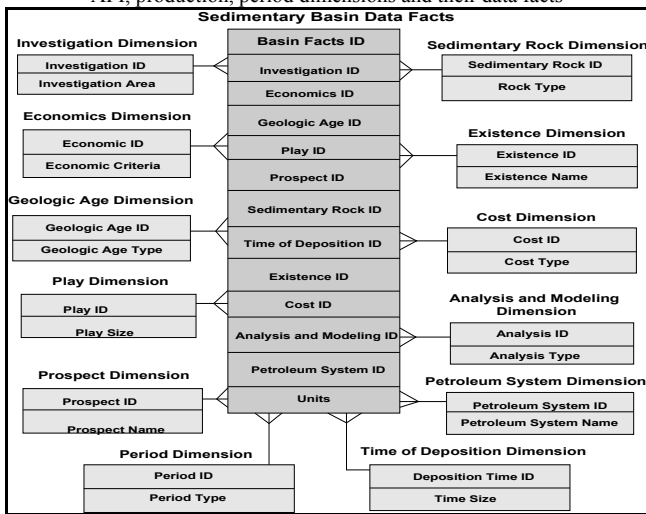

Fig. 5: A star-schema model representing relationships and their connectivity among various attributes of petroleum and depositional systems
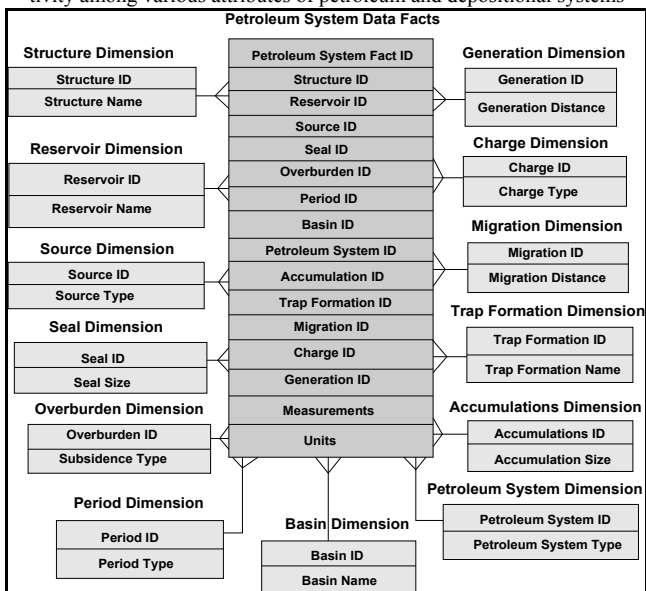
Fig. 6: A star-schema model representing relationships and their connectivity among various attributes of petroleum systems
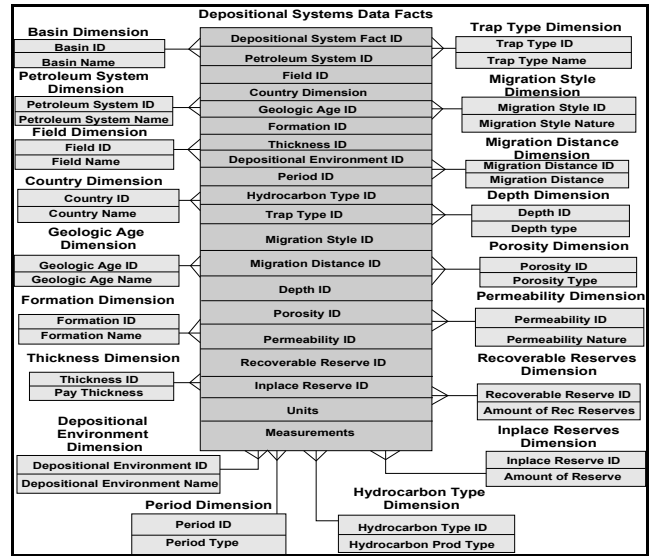

Fig. 7: A star-schema model representing relationships and their connectivity among various attributes of depositional systems of basins

A. Loader Requirements

Because basins and petroleum systems are in large size and their associated data are very large in general. Because of poorly defined syntax and load failures are frequent, could result crashing of loaders. For this reason, DB loaders DB loaders should be able to recover gracefully from errors encountered during parsing of their input files. The loaders are designed to keep loading even in the presence of an error. If partial data has been inserted into the warehouse, a load error maintained on the related dimensions is updated to indicate that an error occurred while parsing the dimension and that the warehouse entry for the dimension may therefore be incomplete or contain errors.

B. Petroleum Systems Warehouse Design

Warehouse schema is designed by first studying the schemas of each DB or super-type dimension to be integrated ([10] and [11]), as well as the schemas of other DBs that use the same data type and semantics. The development of warehouse is guided by several principles DBs schemas for sub-sets, each having similar data type or property.

Since DBs typically conceptualize ingredients of petroleum systems in different ways, any kind of cross-DB operation faces the problem of semantic heterogeneity, whereby, information partitioned in different fields use different definitions (such as different units of measure). One possible approach to supporting petroleum system DBs within warehouse would be to create different schema definitions for each of the conceptualization (domain ontology) of petroleum systems used by the source DBs. Though this approach may achieve the semantic heterogeneity among these DBs, but may be designed not complicating the overall resulting integrated schema of the petroleum systems Warehouse. As narrated in Fig. 8, an integrated framework is designed, explaining data structuring and integration process.
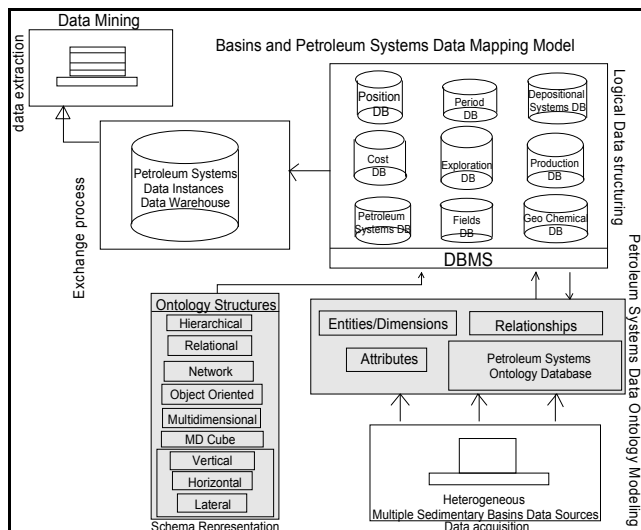
Fig. 8: An integrated framework, modeling data warehouse and mining the petroleum systems

A typical hierarchical structuring may be useful in deducing relationships not only connecting different data dimensions within a petroleum system, but also among different petroleum systems. By and large, this warehouse may typically contain a different petroleum system for every petroleum system DB that it loads. With the result, the warehouse schema could be larger and more complex, and users would have more difficulty understanding the schema, making it more difficult for a user to query the DB, in which case, a user may want to query all petroleum systems in the DB in the warehouse, and essentially write a separate sub-query for each sub-schema.
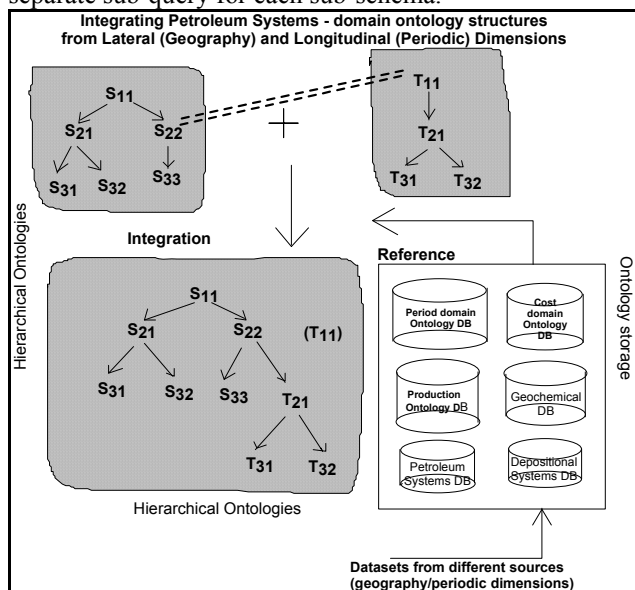

Fig. 9: A hierarchical data structural model

In our approach, instead, a single set of schema definitions covers a given data type, even if that data type is ontologically conceptualized differently in different DBs. For example, a single set of schema definitions is created to span all attributes for petroleum systems that possess similar structural and reservoir characteristics of producing horizons. Another example may be a single set of schema definitions for petroleum systems that have similar geochemical compositions of hydrocarbons. The DB loaders are responsible for translating from the ontology conceptualization (specifications), used in each DB within the family to the domain conceptualization used by the warehouse.

This approach eliminates the semantic heterogeneity of these multiple DBs, allowing users to query all petroleum systems DBs using the same schema. This approach ensures the encoding of the dataset from which each data dimension within the warehouse is derived. Since entries from any petroleum system DBs are loaded into the same set of tables (multi-dimensional schema definitions), it is critical for user queries (data views for interpretation) to be able to distinguish different petroleum bearing basins in the Middle East, or basins in the western hemisphere or eastern hemisphere. Thus queries are made to extract data views for all warehouse shallow marine and or deltaic depositional system DBs that were loaded.

V. RESULTS AND DISCUSSIONS

Multidimensional star schema definitions are integrated in a single schema in a petroleum system warehouse. This type of multidimensional data warehouse respects and maintains the integrity of individual datasets, preserving information about the source of warehouse entries, so that users know and determine exactly, which depositional systems are part of particular set of petroleum systems. Different data views are provided about the same depositional system in different petroleum systems and from different basins. Ontology based multidimensional approach removes the semantic heterogeneity among petroleum systems and basins, removing redundancy among different datasets. This approach also allows detecting the non-redundant petroleum system DBs, for which different algorithms are written for mining purposes. Non-redundant petroleum system DB may be created as a separate schema and DB within warehouse, which can apply to a specific mining algorithmic application. This could satisfy the users, who want to study large size non-redundant petroleum systems DBs.

Schema Implementation and Data Mining Approaches

Petroleum systems warehouse supports several petroleum ecosystem informatics data types, each of which is implemented as one or more tables in the schemas. Dimensional and factual data tables are not shown in the paper, for privacy and protecting the intellectual property. Every dimension in the warehouse has multiple rows in the entry table that defines the metadata such as the time it is inserted in the warehouse, and its time of last update. Every warehouse entity is also associated with the dataset DB from which domain it was derived. Different data views are drawn from a metadata of a warehouse for interpreting trends and correlations [12] that make up the connectivity and communication among relevant data attributes of multiple dimensions. Using grapher and surfer solutions, bubble plots and surface map views are plotted and shown in Figs.11-18. Data relationships plotted between period and other conceptualized (logical) data dimensions have been deduced with interesting connectivity among data properties.

Structure, reservoir, production data attributes, among different petroleum systems and among different depositional systems, provide valuable data correlations and trends among periodically and geographically distributed petroleum systems.
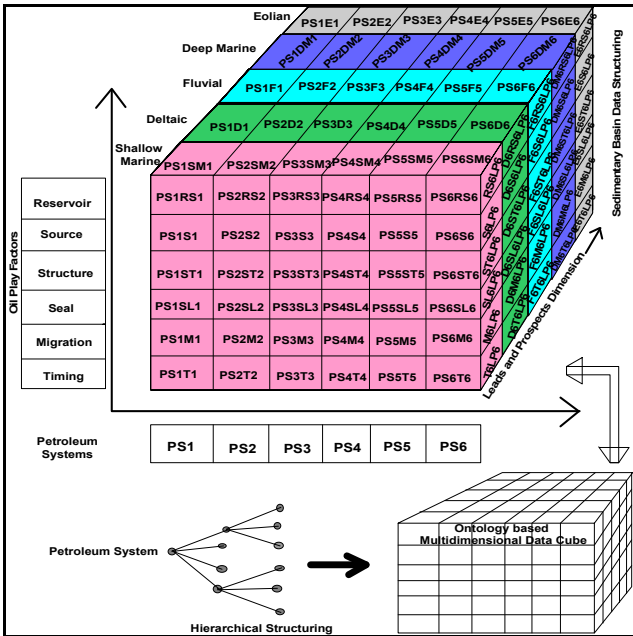
Fig. 10: Multidimensional representation of organizing petroleum systems data in a data cube
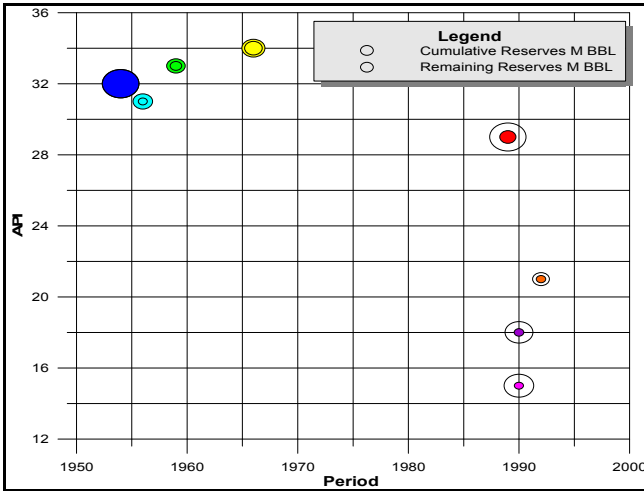

Fig. 13: Deltaic Reservoirs, showing northern hemisphere to southern

Fig. 11: A data view showing API and Cumulative Production with periodic dimension
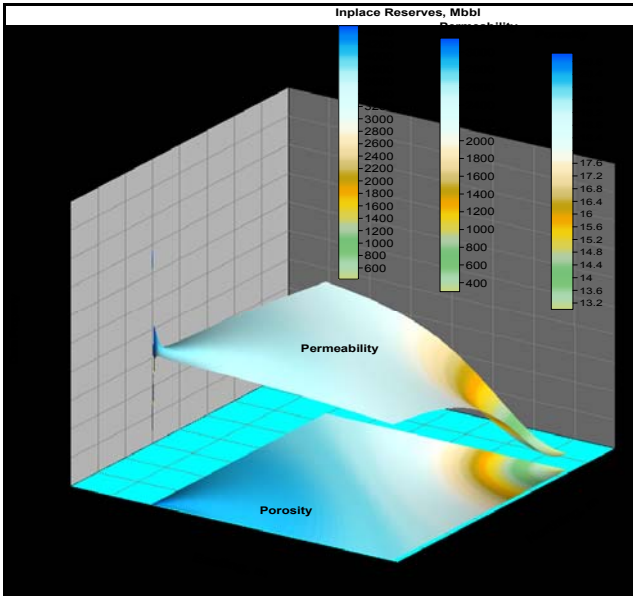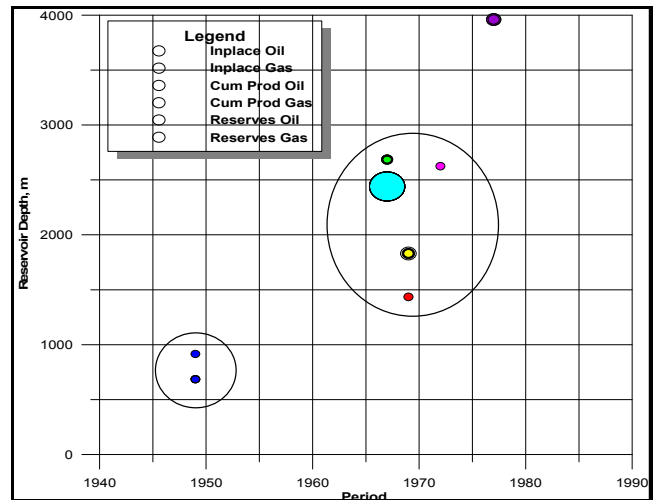

Fig. 14: A data view showing periodic reservoir depths updates

Fig. 12: Shallow Marine Reservoirs, showing from northern hemisphere to southern hemisphere
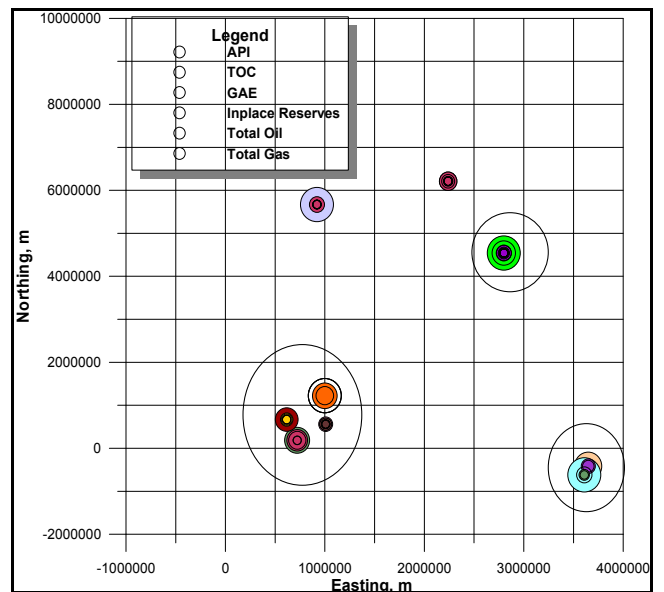

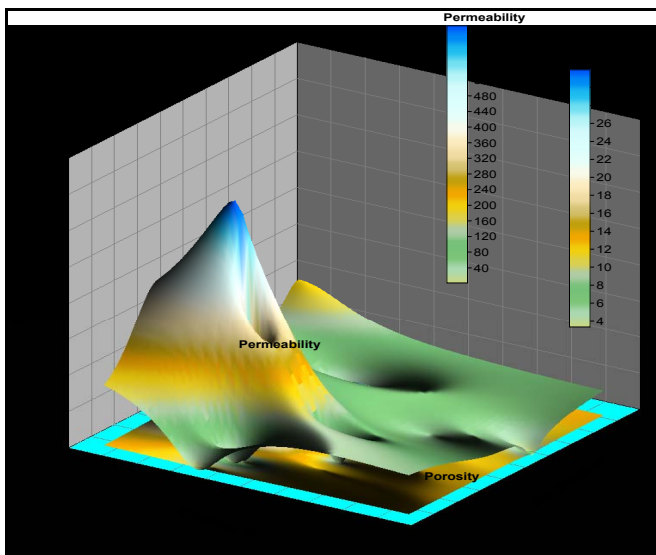Fig. 15: A data view showing source rock potential of petroleum systems

Fig. 16: Surface map view of reservoir properties of world class petroleum systems
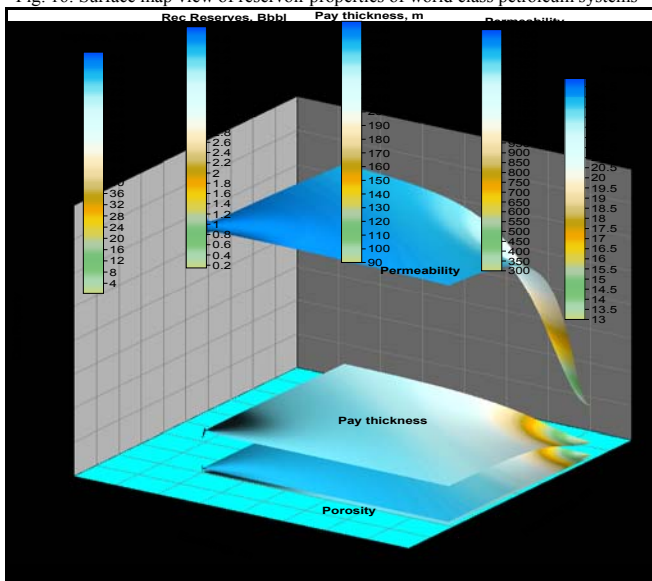

Fig.17: Surface map view of world class reservoir systems
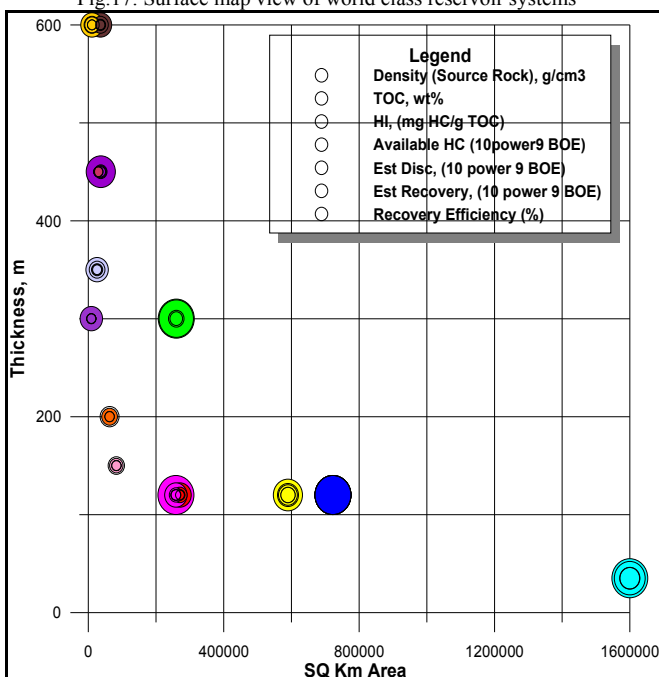

Fig. 18: Multidimensional Data, showing linearity property

VI. CONCLUSION AND SCOPE

• Data warehouse provides a single information repository for massive collection of evaluation of data for Petroleum Systems Analysis and Sedimentary Basin Modeling from different basins. This study will be able to organize and deliver data to different users, who explore and mine the data and perform analysis and interpret different geological events. This will facilitate the decision makers, who make drillable exploratory and development targets in different basins.

• Multiple dimensional data structuring is more flexible with models reusable in different domain ontologies

• Data warehousing approach is more suitable, when considering large size sedimentary basins that comprise of multiple petroleum systems.

• Several data properties are deduced from different depositional systems

• Data attributes and properties drawn among several dimensions show significant correlations, trends and relationships among multiple data attributes such as, structure, reservoir, production, and other geological, geophysical and geochemical data dimensions.

• Present studies suggest interesting correlations among petroleum systems under different depositional regimes.

• Studies are progress in terms of building models for fluvial and deep marine petroleum systems and interpretation of their data views, extracted in our present study.

VII. REFERENCES

[1] Magoom B. L and Dow, D.W (1994), The Petroleum System – from Source to Trap, AAPG Memoir 60, p. 1-625

[2] Parasnis, D.S, Principles of Applied Geophysics, Chapman & Hall, 1997

[3] Thomas, J.L, Yannick, P. Valerie, W., Gupta, P. Stringer-Calvert, D.WJ, Tenenbaum, J.D and Karp, P.DBiowarehouse: a bioinformatics database warehouse toolkit; BMC Bioinformatics 2006, 7:170, p.1-14; http:///www.biomedcentral.com/1471-2105/7/170

[4] Hoffer, J.A, Presscot, M.B and McFadden, F.R, Modern Database Management, 7th Edition, Prentice Hal l, 2005

[5] Pujari, A.K. Data Mining Techniques, 2002, University Press, 2002, p. 7-66

[6] Davidson SB, Overton C, Buneman P: Challenges in integrating biological data sources, Journal of Computational Biology 1995, 2(4): 557-572

[7] Karp P: A strategy for database interoperation. Journal of Computational Biology 1995, 2[4]: 573-586

8] Nimmagadda, S.L. and Dreher, H. (2006) Ontology-base data warehousing and mining approaches in petroleum industries: in Negro, H.O, Cisaro, S.G. and Xodo, D. (Eds.), Data Mining with Ontologies: Implementation, Findings and Framework, a book published in 2007 by Idea Group Inc. http://www.exa.unicen.edu.au/dmontolo/

[9] Nimmagadda, S.L, and Dreher, H. (2006a) Mapping and modelling of Oil and Gas Relational Data Objects for Warehouse Development and Efficient Data Mining, a paper presented and published in the *proceedings of the 4th International Conference of IEEE Industry Informatics*, held in Singapore, August

10] Nimmagadda, S.L. and Dreher, H., Rudra, A (2005a) Ontology of Western Australian petroleum exploration data for effective data warehouse design and data mining, *Proceedings of 3rd international IEEE conference on Industrial Informatics*, held in Perth, Australia

[11] Rudra, A and Nimmagadda, S.L. (2005) Roles of multidimensionality and granularity in data mining of warehoused Australian resources data, Proceedings of 38th Hawaii International Conference on Information Sciences, Hawaii, USA

[12] Shawkat Ali, A. B. M. and Wasimi, S. A. (2007) Data Mining: Methods and Techniques, p. 196-219 and p. 25-267